DOCUMENT RESUME

ED 349 809 FL 020 588

AUTHOR	Stansfield, Charles M.; And Others
TITLE	The Preliminary Chinese Proficiency Test (Pre-CPT): Development, Scaling, and Equating to the Chinese Proficiency Test (CPT). Technical Report 1.
INSTITUTION	Center for Applied Linguistics, Washington, D.C.
SPONS AGENCY	Office of International Education (ED), Washington, DC.
PUB DATE	92
CONTRACT	P017A00001
NOTE	67p.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Chinese; Higher Education; *Language Proficiency; *Language Tests; National Norms; Secondary Education; *Second Language Learning; *Test Construction; *Test Format; Test Items; Test Reliability; Test Validity
IDENTIFIERS	*Preliminary Chinese Proficiency Test

ABSTRACT

This report describes the development, construction, and validation of the Preliminary Chinese Proficiency Test (Pre-CPT), a standardized, nationally-normed test of listening and reading comprehension for beginning-level native English-speaking learners of Chinese as a second language. The Pre-CPT was designed as a lower-level version of the commonly-used Chinese Proficiency Test (CPT). It is intended to accompany the CPT and Pre-CPT combined test interpretation manual. The report's eight chapters describe: the project's background, test structure and content, test administration time, and test materials; the initial phase of the test development process, focusing on committee membership and deliberations; development of the field test form of the Pre-CPT and the subsequent field testing; the nationwide norming administration in Chinese language programs nationwide; construction of a common scale for the Pre-CPT and CPT; the final form of the Pre-CPT and final norming means and item difficulty results; and psychometric properties of the test (reliability, precision of measurement, validity, test subscore intercorrelations). A brief list of references is included. Appended materials include lists of field tests and norming participants, a field testing examinee background questionnaire, norming tables for the Pre-CPT and CPT, and sources for test content. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

CAL

Technical
Report 1

Summer 1992

ED349809

The Preliminary Chinese Proficiency Test (Pre-CPT): Development, Scaling and Equating to the Chinese Proficiency Test (CPT)

Charles W. Stansfield
Dorry Mann Kenyon
Xixiang Jiang

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G.R. Tucker
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Center for Applied Linguistics
Washington, DC

FL 020 588

BEST COPY AVAILABLE

Acknowledgements

The authors would like to thank the many test supervisors and students who participated in the field testing and the norming administration of the Pre-CPT. Their participation in this test development project was invaluable.

Thanks are also due to Jianhua Bai (Kenyon College), Fred Davidson (University of Illinois), and Huixing Tang (The Psychological Corporation) for their helpful comments on an earlier version of this report.

We also acknowledge the many contributions to this project made by the following staff members of the Division of Foreign Language Education and Testing at the Center for Applied Linguistics: Kenneth Martens Friesen, Margaret Malone, Lisa Pinson, Sylvia Rasi and Laurel Winston.

Copyright © 1992 by Center for Applied Linguistics, Washington, DC. All rights reserved.
Based on the Final Report for Grant No. PO17A00001, Office of International Research and Studies, U.S. Department of Education

Preface

This technical report is designed to serve as an accessory document to the combined *CPT and Pre-CPT Combined Test Interpretation Manual*. The *Combined Test Interpretation Manual* provides basic information to the test score user so that he or she can interpret scores on either test. However, there exists a need for a more detailed document to provide a greater understanding of the test. Language testing researchers, language test developers, language educators, and others may have a number of questions that it is not feasible to address in the *Combined Test Interpretation Manual* because of the audience the *Manual* is intended to serve. It is hoped that this technical report will answer the questions those individuals might have.

The report is also designed to serve as a stand-alone document, at least in the most minimal sense. Thus, it includes a basic description of the test, and it details the decisions that were made and the procedures that were followed during test development, field testing, norming, scaling and equating.

On the other hand, the report is not intended to replace other test program publications, such as the *CPT/Pre-CPT Combined Test Interpretation Manual* and the *CPT/Pre-CPT Combined Examinee Handbook*. Thus, it does not contain sample items, nor does it describe program policies and procedures. Only by reading all of these publications can one gain a complete understanding of the Pre-CPT and the CPT.

CAL is pleased to make this report available to the field. The report is an accurate description of the process of how a particular test was developed. Since the project broke new ground in a number of respects, to some extent we learned as we progressed. This was particularly true of the equating process. We hope others can learn from our experience, and we offer this detailed report on the project to that end.

TABLE OF CONTENTS

Preface i

1. Introduction 1
   1.1 Background 1
   1.2 Structure of the Pre-CPT 2
   1.3 Content of the Pre-CPT 4
   1.4 Test Administration Time 5
   1.5 Pre-CPT Test Materials 6

2. Project Start-Up 7
   2.1 Test Development Committees 7
   2.2 Initial Committee Meetings 9
      2.2.1 Initial Local Advisory Committee Meeting 9
      2.2.2 Initial External Advisory Committee Meeting 9
      2.2.3 Initial Item Writers Committee Meeting 10

3. Development of the Field Test Form 11
   3.1 Initial Development of the Test Items 11
   3.2 Review of the Test Items by the Local Advisory Committee 12
   3.3 Development of the Preliminary Test Form 12
   3.4 Preparing the Field Test Materials 12

4. Field Testing 15
   4.1 Administration of the Test 15
   4.2 Results of the Field Testing 16
   4.3 Qualitative Input 19
   4.4 Revision of the Test 20

5. The Norming Administration of the Pre-CPT 21
   5.1 Rationale for the Norming Administration 21
   5.2 The Results of the Norming Administration 22

6. Building a Common Scale for the Pre-CPT and the CPT 25
   6.1 Concurrent Calibration of the Pre-CPT and the CPT 26
   6.2 Building the CPT Scale 29
   6.3 Building the Norming Tables 30

7. The Final Form of the Pre-CPT 33
   7.1 The Selection of Items 33
   7.2 The Final Norming Tables 34
   7.3 Difficulty of the Test 37

8. Psychometric Properties of the Pre-CPT and CPT 39
8.1 Reliability 39
8.2 Precision of Measurement 40
8.3 Validity 41
8.4 Intercorrelations Among Test Subscores 42

REFERENCES 45

List of Appendices

Appendix A: Pre-CPT Field Test Participants
Appendix B: Examinee Background Questionnaire
Appendix C: Pre-CPT Norming Administration Participants
Appendix D: Norming Tables for the Pre-CPT and CPT
Appendix E: Standard Error of the Estimate for the Pre-CPT and CPT Across CPT Scale Scores
Appendix F: Sources for Pre-CPT Listening and Reading Passages

1. Introduction

This report describes the development of the *Preliminary Chinese Proficiency Test* (Pre-CPT), a standardized, nationally-normed test of listening and reading comprehension for beginning level English-speaking learners of Chinese. The Pre-CPT was developed by the Center for Applied Linguistics (CAL) under the auspices of the United States Department of Education (Grant No. PO17A00001) with the cooperation of numerous linguists and Chinese language experts from a variety of academic institutions across the country. The project was initiated in September, 1990 and completed in June, 1991. The test was developed in response to requests from the Chinese teaching profession for a lower-level version of CAL's successful *Chinese Proficiency Test* (CPT), which has been used by more than one hundred institutions. Many of these institutions use the CPT on a regular basis.

This introductory chapter provides the background to the development of the Pre-CPT, beginning with a description of the relationship between the CPT and Pre-CPT and concluding with the description of the structure, format and content of the Pre-CPT.

1.1 Background

The original *Chinese Proficiency Test* (CPT) was developed in 1984 by the Center for Applied Linguistics, in close collaboration with representatives of the Chinese language teaching profession in the United States. Funding was provided by the U.S. Department of Education. The test was produced to meet the need for an objective measurement of Chinese language proficiency. It was designed to evaluate the level of general proficiency in Chinese listening and reading comprehension attained by English-speaking learners of Chinese. This test was the first professionally-developed, standardized proficiency test in the Chinese language teaching field based on real-world language use rather than on textbook language. In its development, the CPT followed the reading and listening proficiency guidelines established by the Federal Interagency Language Roundtable (FILR) and American Council on the Teaching of Foreign Languages (ACTFL). The reading and listening stimuli were taken from real-life language tasks identified in FILR levels 1, 2 and 3 (equivalent to ACTFL levels Intermediate, Advanced and Superior).

The CPT consists of two sections: Listening Comprehension and Reading Comprehension. The Reading Comprehension section contains two sub-sections: Structure and Reading. The test has a total of 150 4-option multiple choice items, with 60 items in Listening Comprehension, 35 items in Structure and 55 in Reading. The total testing time is two hours. The test is machine-scored and examinees are provided with scores for the three individual parts, as well as a total score. Norms are published by CAL to help score users interpret examinee scores.

Over the years, the CPT has received wide acceptance and favorable comments from the Chinese language teaching profession. Since its development, the test has been used for a variety of purposes. These include:

- admission to and placement within Chinese study programs
- exemption from Chinese language requirements
- applications for scholarships
- competency testing upon exit from Chinese programs
- measurement of progress during the course of instruction

The test has been administered to more than 2500 examinees by more than 110 different institutions, many of which are regular users of the CPT.

The need for a second form of the CPT arose with the increased use of the test. The CPT office at CAL continually received requests for a second form of the test. Among these requests, the most acute was the request for a lower-level version of the test. Since most items on the current CPT are designed for students at the Advanced and Superior levels of the ACTFL scale, it was felt that there should be a lower-level CPT, focused on the Novice and Intermediate levels on the ACTFL scale, to serve the needs of the large number of students beginning college and high school Chinese language programs.

The Division of Foreign Language Education and Testing at the Center for Applied Linguistics proposed the development of such a Chinese language proficiency test to the Department of Education. The proposal was accepted and funded, and a year's cooperation between test development specialists and Chinese language experts has resulted in this report and the Pre-CPT program.

1.2 Structure of the Pre-CPT

Paralleling the CPT, the Pre-CPT has three sections: Listening Comprehension, Reading Comprehension and Structure. All items on the Pre-CPT are four-option multiple choice.

In the Listening Comprehension section, examinees hear the stimulus in Chinese, followed by a question about it in English. Both the stimuli and the question are heard from a professionally-recorded Master Tape. After hearing the English question, examinees then choose one of the four answers printed in the Pre-CPT Test Booklet and mark their choice on the machine-readable Pre-CPT Answer Sheet.

In the Reading Comprehension section, students are presented with a variety of Chinese texts, each followed by a question about it in English. The examinee is provided with four response choices. The Chinese text, the English question and the four choices are all printed in the Pre-CPT Test Booklet. Chinese reading passages are printed in both traditional and simplified character forms. The content of both forms is identical, and examinees may refer to either or both of the forms while working through the sections.

In the Structure section, examinees are presented with five written Chinese passages that each have five missing portions in them. For each missing portion, examinees are presented with four suggested completions. Examinees are required to choose one of them and mark their answer

on the answer sheet. In this section, all Chinese materials are presented side-by-side in both traditional and simplified character forms. Again, examinees may refer to either or both of the forms while working through this section.

Although there are great similarities in the organization of the two tests, the Pre-CPT is organized slightly differently from the CPT. Whereas the CPT has two sections, Listening Comprehension and Reading Comprehension, with Reading sub-divided into Structure and Reading Comprehension, the Pre-CPT has three separate sections: Listening Comprehension, Reading Comprehension, and Structure. In the CPT, administration of the Structure section is mandatory, while in the Pre-CPT, it is optional. This design is intended to make the Pre-CPT more flexible and more appropriate for programs that do not stress grammar in beginning level Chinese instruction. If desired, the three sections of the Pre-CPT can be administered in two separate sessions.

Table 1 provides an overview of the organization of the Pre-CPT.

Table 1
Organization of the Pre-CPT

Section	Total Time	Number of Items	Part	Format of Stimulus	Number of Items
Listening Comprehension	25 min (approx)	50			
			One	Utterances	20
			Two	Dialogues	20
			Three	Monologues	10
Reading Comprehension	45 min	50			
				Nonlinear Text	12
				Signs	8
				Passages	30
Structure	20 min	25		Paragraphs	

Although in general items on the Pre-CPT are arranged in order of increasing difficulty, they are also grouped according to the format of the item stimulus. For example, items in the Listening Comprehension section are grouped in three parts according to the format of the listening passage. Items in the first part, "utterances," present the examinee with the speech of a single speaker excerpted from a longer dialogue or conversation. Items in the second part present short, intact dialogues between two speakers. Items in the third part present examinees with speech found in naturally-occurring monologues, such as those from radio announcements or news broadcasts.

Although there are no separately designated parts in the Reading Comprehension section, these items are also grouped by the format of the reading stimulus. There are three groupings in this section. The first is "Nonlinear Text," which refers to reading stimuli that are non-prose

forms of writing existing outside the normal conventions of sentence, paragraph and text structure. Examples include the writing found in book titles, on stamps, in schedules, on identification cards and the like. The second grouping consists of the eight "Sign" items. These items, also nonlinear text, present characters found on signs commonly seen in China and Taiwan. Items in the third grouping, "Passages," present examinees with prose and expository text following normal conventions of sentence and paragraph structure.

Items in the Structure section are of a cloze (deletion) type focussing on aspects of grammar and syntax. These items consist of deletions in paragraphs, with five deletions per paragraph.

1.3 Content of the Pre-CPT

The development of the Pre-CPT followed strict guidelines in terms of the test content covered. To the greatest extent possible, all Chinese materials on the Pre-CPT are drawn from authentic sources; i.e., Chinese language materials prepared for Chinese native speakers. Sources for listening passages included Chinese movies, recordings of conversations, and radio and television programs. For reading and structure passages, sources included Chinese magazines, journals, newspapers, books, movie scripts, informational brochures, medical prescriptions, and a collection of public signs and personal documents.

Every effort was made to include as wide a range as possible of the various social and institutional interactions that would most likely be encountered in real-life language-use situations in a Mandarin Chinese-speaking environment. The content covered on the Pre-CPT can be described in terms of topic areas (subject of the listening or reading passage) and language functions (speaking tasks that are covered in the listening passages) or language text types (of reading passage material). Table 2 presents an overview of the content of the Pre-CPT by section of the test. The first column lists the topic areas covered, the second lists the language functions (for listening passages) or the language texts (for reading and structure passages) found in the items, and the last column lists the sources of listening and reading passages used as test stimuli.

Table 2
Overview of the Content Covered on the Pre-CPT

Section	Topic Areas	Language Functions/Text Types	Sources
Listening	education family food health media social life sports transportation	advising apologizing comforting comparing complaining giving instruction informing leave-taking making an announcement making an appointment making a comment making an introduction making a purchase making a request offering stating a preference	movies radio shows recorded conversations speeches telephone messages television
Reading	art consumerism education family government health media social life transportation	advertisements bulletins captions instructions interviews labels letters narrative prose notes personal IDs scripts schedules signs	journals magazines movie scripts newspapers novels street signs
Structure	education family travel biography sports	descriptive prose narrative prose	brochures newspapers texts

1.4 Test Administration Time

The three sections of the Pre-CPT may be administered in either one or two sessions, with the administration of the Structure section being optional. In a two-day session, the Listening

Comprehension section and the Structure section (if included) are administered at the first session, and the Reading section administered at the second. Exclusive of passing out test booklets, filling in information on the machine-readable answer sheet, and taking care of other administrative matters, the Listening Comprehension section requires approximately 25 minutes, the Reading Comprehension section 45 minutes, and the Structure section, 20 minutes. If given in one sitting, the total testing time is about one hour and a half, though the whole administration may require just under two hours. If administrative matters are attended to before the testing session begins, the Pre-CPT can be administered in two 50 minute class periods. Two 55 minute periods are sufficient for handling the complete administration of the Pre-CPT.

1.5 Pre-CPT Test Materials

The operational program of the Pre-CPT consists of the following materials:

- *Pre-CPT Test Booklet*, which contains all test instructions, the options for the items of the Listening Comprehension section, the Chinese text, English questions and options for the Reading Comprehension section, and the Chinese text and completion options for the Structure section

- *Pre-CPT Master Tape*, which is a professionally recorded audio tape containing the listening stimuli and the English questions for the Listening Comprehension section of the Pre-CPT

- *Pre-CPT Answer Sheet*

- *Pre-CPT Supervisor's Manual*, which contains all the necessary instructions for administering the test

- *CPT/Pre-CPT Combined Examinee Handbook*, covering both the Pre-CPT and the CPT, which familiarizes examinees with both tests in order for them to be ready to take either one, reviews basic information about the tests and gives sample test items for both.

- *CPT/Pre-CPT Combined Test Interpretation Manual*, covering both the Pre-CPT and the CPT, which helps test users interpret test scores so that they can be meaningfully used in their programs and presents a brief description of both tests and guidelines for deciding which is most appropriate for a specific program.

2. Project Start-Up

The initial phase of the Pre-CPT project involved a great many efforts to ensure a sound theoretical and methodological framework for the development of the test items.

This chapter describes the preparation process of the test development project. It encompasses both logistic issues and theoretical concerns.

2.1 Test Development Committees

The day-to-day activities of the project were coordinated and directed by a team of CAL staff members. Charles W. Stansfield served as the Director of the Project, with Dorry Mann Kenyon serving as Assistant Project Director and Xixiang Jiang as Project Coordinator.

To assist in the development project, three test development committees were formed. The first was the External Review Committee, consisting of Chinese linguists and Chinese language experts from across the nation. Many members of this committee served on the development committee of the original CPT. Below is a list of the members of this committee.

External Advisory Committee

Jianhua Bai	Kenyon College
Telee Richard Chi	University of Utah
Albert E. Dien*	Stanford University
Ying-che Li*	University of Hawaii
Timothy Light*	Middlebury College
Shou-hsing Teng*	University of Massachusetts
Galal Walker*	Ohio State University

* Members of the original CPT test development committee

Members of this committee were asked to review specifications for the test, review the test forms before pilot testing, and to review revisions made to the test forms before the norming administration. Whenever possible, these committee members also helped coordinate pilot testing of the Pre-CPT at their respective institutions.

The second committee was the Local Advisory Committee. Members of this committee were Chinese language professors, experienced language instructors and high school Chinese teachers resident in the Washington, DC area. Members of this committee are listed below.

Local Advisory Committee

Neil Kubler	Williams College
Davis Lee	George Washington University
Hung Yi Shen	University of Maryland
Wayne Smith	Foreign Service Institute
Richard Thompson	Georgetown University
Ronald Walton*	National Foreign Language Center
Gwen T. Wang	Richard Montgomery High School

* Member of the original CPT test development committee

Members of the local advisory committee met to draft the initial test specifications, reviewed items under development, reviewed the test form before pilot testing and the norming administration, and served as consultants when special problems arose. They also helped make arrangements for pilot testing the Pre-CPT.

The third group, the Item Writing Committee, was composed of experienced teachers from local universities and high schools. These were:

Item Writers Committee

Yuling Pan	Diplomatic Language Services
Lina Xie	Sidwell Friends High School
Hannah Wu	Bell Multicultural High School
Weiping Wu	Georgetown University

The item writers were trained by CAL staff and met together with CAL staff on a weekly basis between October and December, 1990. They were responsible for finding suitable listening and reading passages from authentic sources and drafting items to test listening and reading comprehension.

In addition to these three working committees, broadcast professionals from television and radio institutions were involved in recording the Pre-CPT Master Tape: Helen Shen and Dong C. Wang, from the Voice of America, and Tong Shen and Caroline C. Wang, from Channel 56, a local television channel with Chinese language programming.

Given their various areas of expertise and experience in teaching and testing, and in the Chinese language as used both in mainland China and elsewhere, members of all committees contributed to the success of the Pre-CPT project.

2.2 Initial Committee Meetings

Each committee met within two months of the project start-up. The results of their initial meetings, which set the course of the project, are described below.

2.2.1 Initial Local Advisory Committee Meeting

The Local Advisory Committee met in early October, 1990, to develop the test specifications, to set down the guidelines with which the test item writers would work, and to discuss issues relating to the development of the test items. Several theoretical concerns were addressed at the meeting regarding concepts of proficiency, authenticity, and the target level for the test. Members of the Committee conferred for a day and a half and concluded with a general agreement on the issues considered.

It was agreed that this test, being a proficiency test, was to test examinees' ability to function in an authentic Chinese-speaking environment. Therefore, it would not be designed to accommodate any specific Chinese language teaching curriculum, nor would any sort of achievement test be appended to this proficiency test. It was also acknowledged that the materials used in the items should be authentic; i.e., language materials that are produced by native speakers for use by native speakers in their native environment. The decision to use authentic materials as stimuli was made in order to promote a trend in the language teaching field toward increased use of real language to achieve communicative competence. It was also recognized that the target level of the test should be between O+ and 1+ on the FILR scale, which corresponds to the levels of Novice Mid/High to Intermediate High on the ACTFL scale. In relation to the number of contact hours of instruction received, this level of proficiency was thought to translate into approximately one year of college level instruction or three years of high school instruction.

In terms of the test item format, it was agreed at the meeting that four-option multiple choice formats should be used throughout the test, in conformity with the demands of large scale testing. It was also agreed that a cloze format be added in the Reading Comprehension section to test both structural and lexical knowledge within the context of extended discourse. As for the length of the test, it was suggested that, since the Pre-CPT is to be at a lower level, the time required to take the test should be shorter than for the CPT.

2.2.2 Initial External Advisory Committee Meeting

Members of the External Advisory Committee were sent copies of the minutes of the Local Advisory Committee Meeting for review and comments. In addition, members were invited to a meeting during the annual ACTFL and Chinese Language Teachers Association (CLTA) conference on November 19, 1990. Among the issues discussed at this meeting were sources of authentic but simple language materials appropriate for the target level of the test and specific methods for conducting the analysis of test items after pilot testing. The Committee advised CAL staff that easy but authentic written language material in Chinese could be found in some movie transcripts and short novels.

The members of the External Advisory Committee reached a consensus regarding the practicality and suitability of the test specifications and overall guidelines recorded in the minutes of the Local Advisory Committee Meeting. They also expressed their willingness to offer any assistance needed for the project.

2.2.3 Initial Item Writers Committee Meeting

CAL staff designed and conducted a two day intensive training program for the four members of the Item Writers Committee. The training program provided the item writers with an opportunity to familiarize themselves with the test specifications and general guidelines recommended by the advisory committees for the project. Charles W. Stansfield, Project Director, instructed the group on writing cloze-type items and items to test reading comprehension. Huei-ling Worthy, a government language school instructor and an ACTFL-certified oral proficiency interviewer, was invited to explain the ACTFL Chinese Proficiency Guidelines to the item writers and to instruct them on developing items to test listening comprehension.

Practice in item writing was conducted towards the end of the training session. This training enabled the item writers to apply their knowledge of the Chinese language and experience in teaching to developing test items according to the stipulations set forth for this specific project. Since the item writers were teachers from both mainland China and Taiwan, their work and collective revision on items under development ensured a balance between language forms used on the mainland and those used in other areas where Mandarin Chinese is spoken.

3. Development of the Field Test Form

After the initial start-up of the project, attention was focused on the development of Pre-CPT test form for field testing. This chapter describes the steps in developing the field test form.

3.1 Initial Development of the Test Items

The four members of the Item Writers Committee were divided into two groups of two. One group focused on developing listening items, the other on reading items. Structure items were prepared by the project coordinator and one of the item writers. From October to December, item writers worked on items at home and attended weekly meetings with CAL project staff. The purpose of these meetings was to provide further training in item writing; to review, critique, and revise items under development; and to ensure that work proceeded on schedule.

Following instructions by CAL staff, item writers first identified authentic listening or reading passages suitable for testing comprehension. They then determined an appropriate aspect of each passage to test. They then wrote the question, chose appropriate distractors, and submitted the items for review at the weekly meetings.

To help item writers focus on the task of developing quality items, each item was submitted on a form that required item writers do several things. First, they documented the source of the passage and indicate its content area and topic. For the language used in the passage, they separately indicated whether vocabulary and grammar were, in their opinion, of low, medium, or high difficulty, and whether there was strong, medium or weak contextual support. The item submission form also encouraged item writers to carefully consider various aspects of the question they were writing for the passage. Item writers indicated on the form 1) the **cognitive task** involved in determining the correct answer (that is, whether to show understanding of basic learned material, the main idea of the passage, facts/details mentioned in the passage, or an inference based on the passage) and 2) the **relative importance of knowledge of vocabulary, grammar, understanding of contextual clues or pragmatics** in determining the correct answer. Finally, in order to help item writers remain conscious of the item's overall difficulty level, the form asked them to indicate the intended difficulty level on the ACTFL scale.

After items were revised by the item writers themselves, following review and input at the weekly meetings, they were then reviewed by CAL project staff and either accepted, discarded or returned to the original item writer for further revision. Accepted items were put into the Pre-CPT item bank. Regular follow-ups were conducted to ensure that the items being accepted were fulfilling the specifications for the test.

3.2 Review of the Test Items by the Local Advisory Committee

When the item bank contained more items than the number required by the test specifications, items were then sent to the members of the Local Advisory Committee for review. Each committee member was assigned a subset of the items, all from one section of the test. All items were reviewed by at least three committee members. On the basis of input received, items were either kept as they were, revised as per specific comments, or removed from the item bank.

3.3 Development of the Preliminary Test Form

From the items remaining in the item bank, CAL staff assembled the Preliminary Test Form, following guidelines set forth in the test specifications. Recognizing that items might be rejected after field testing, this form contained more items than were envisioned for the final form: 100 listening comprehension items, 80 reading comprehension items, and 30 structure items. The draft test booklet and listening script were prepared and sent to the members of the External Review Board for comment.

After the draft test form was revised on the basis of input from members of the External Review Board, the field test materials were prepared.

3.4 Preparing the Field Test Materials

The Pre-CPT test booklet for the field testing was prepared using both WordPerfect 5.0 (for English-only sections) and BrushWriter, a Chinese word processing program obtained for this project, for sections containing Chinese characters or Chinese and English mixed. Brushwriter allowed for the printing of both traditional and simplified forms of characters side-by-side in the text. Realia used in the test, such as stamps, identification cards, and diplomas, were photo-copied and then inserted into the test booklet.

To ensure that all of the listening passages were clear and of professional quality, they were re-recorded using professional broadcasters from local Chinese television and radio stations. After auditions were conducted, two male and two female voices were selected. These speakers were then instructed on how to ensure a natural delivery of the listening passages. The speakers strove to read the written script as naturally as possible, while keeping in mind the need for clear articulation. During the recording session at a professional studio, the project coordinator, who served as the director of the session, and two other individuals, one an advanced-level student of Chinese, the other a Chinese language teacher, listened critically to each passage as it was recorded. These three persons either agreed that a take was of appropriate quality for inclusion on the test, or made suggestions for another rendition. English sections of the Master Tape (test instructions and questions on the listening passages) were recorded by a professional radio announcer. The recordings were edited by professional staff at the recording studio.

The test booklet used during field testing contained 76 items in the Listening Comprehension section, 64 items in the Reading Comprehension section, and 30 items in the Structure section. In addition to the test booklet and master tape, other materials were prepared for the field testing. These included machine-readable answer sheets, instructions for test administration, written instructions on collecting background information from Pre-CPT field test examinees, and a test familiarization sheet to be given to examinees prior to taking the test. When this process was completed, the Pre-CPT was ready to be field tested.

4. Field Testing

To ensure that the developed Pre-CPT form was valid and appropriate for the target group for which it was designed, the Pre-CPT was field tested during late February and early March of 1991 on examinees from both university and high school Chinese language programs. This chapter describes the field testing procedures and the results.

4.1 Administration of the Test

The Pre-CPT was designed to test students in the range extending from Novice Mid to Intermediate High according to the *ACTFL Proficiency Guidelines* for Chinese (ACTFL, 1987). Since language programs do not classify themselves according to the ACTFL scale but in terms of years and credits, project staff felt that students at the end of the second semester of a first year college level course (meeting at least five hours a week) or the third or fourth year in a high school program should be the target population for field testing. In addition, examinees at both lower and higher ends of this range were to be included (i.e., students in their first, third and fourth semester in college, and students in their second or fifth year of high school Chinese language programs) to compare examinee performance on a broader scale.

With the help of members of both the local and external advisory boards, a number of institutions were invited to participate. A total of 16 institutions, 11 colleges and 5 high schools, took part in the field testing. With the cooperation of the volunteer teachers, the field test version of the Pre-CPT was administered to a total of 299 students between the period of late February and early March, 1991. All students participating in the field testing completed a background questionnaire. The number of examinees completing at least one section of the test from each participating institution or school can be found in Appendix A: Pre-CPT Field Test Participants.

Examinees completed a background questionnaire before taking the field test version of the Pre-CPT (see Appendix B: Examinee Background Questionnaire). Of those participating in the Pre-CPT field test, 36% reported that they are ethnic Chinese. 8.5% reported that they speak Mandarin Chinese at home, while 14.3% reported that they speak a Chinese language other than Mandarin. At the time of the test administration, the participants were enrolled in classes ranging from second year Chinese at the high school level to sixth semester Chinese at the college level. Most of them were currently enrolled either in their third year of high school Chinese (13.6%) or their second (25.7%) or fourth (25%) semester of college-level Chinese. In other words, 64.3% were from these three levels. The majority of the students (65.8%) indicated that they were participating in or had completed one year of Chinese instruction prior to taking the Pre-CPT. From these demographics, it can be seen that a majority of the participants fell into the target group of the Pre-CPT.

A summary of the most important demographic information is presented in Table 3.

Table 3
Demographic Information on
Field Test Participants

Total Number	299
Ethnicity	
Chinese	36%
Non-Chinese	64%
Languages	
Mandarin spoken at home	8.5%
Other Chinese language spoken at home	14.3%
No Chinese spoken at home	77.2%
Level of Chinese Instruction	
3rd or 4th year high school, or 1st year college	64.3%
Other	35.7%

4.2 Results of the Field Testing

Both quantitative and qualitative data were collected on the field test form. Quantitative data consisted of the examinees' responses to the test items. Qualitative data consisted of comments made by test supervisors.

Examinees recorded their responses to the background questionnaire and their answers to the test on an NCS (National Computer Systems) General Purpose Answer Sheet. Each sheet was scanned twice on CAL's NCS Sentry 3000 Optical Scanner: the first time to collect background information using the program *Scantools*, the second to score the test using the *MicroTEST Score II Plus* program. The two databases thus entered were merged into one file using the *Paradox* database system. Item analyses were conducted using the *Test Analysis Program*, a classical item and test analysis program, and statistical analyses were performed using SAS.

A test analysis was first conducted on the entire group of examinees. Table 4 summarizes the descriptive statistics by section.

Table 4
Descriptive Statistics from the Pre-CPT Field Testing

Section	Number of Examinees	Number of Items	Mean Score	Std. Dev.	Reliability	Mean P-value	Std.Dev. P-value
Listening	266	76	52.59	14.20	.96	.69	.15
Reading	254	63	43.72	12.67	.95	.69	.14
Structure	262	30	17.35	7.43	.91	.58	.12

Note: One of the original 64 Reading items was double-keyed and thus excluded from analysis.

The results show that the subtest reliabilities were quite high, which is most due to the length of each section and the fact that a wide range of abilities were represented in the sample. The

mean p-value for the listening and reading items was appropriate for a multiple choice test. However, the Structure section seemed rather difficult for the sample.

The next step was to analyze the individual item statistics to detect if there were any malfunctioning items. *Test Analysis Program* (TAP) gives a wealth of information for studying this, including point-biserial correlations (as a discrimination index), p-values, frequency of responses to each item broken up by quintiles, and graphic representation of the percent of correct responses to test items by quintiles. In terms of discrimination, only items with a point-biserial above .30 were considered acceptable; most for the Listening and Reading Comprehension section were above .45. An analysis of the ability of the distractors to discriminate also revealed that the vast majority of items in the Listening and Reading Comprehension sections were problem-free. Quite a few of the individual cloze items in the Structure section were too difficult, however, and did not discriminate well between examinees at different ability levels.

Since most of the listening and reading items were statistically acceptable, it was necessary to examine the difficulty of the items in order to select properly those for inclusion on the final test form. To do this, it was first necessary to determine the extent to which the performance of the entire group reflected that of the target group of second semester college level and third and fourth year high school students. Items appropriate for the final form should be in a difficulty range appropriate for this latter group. Table 5 shows the mean scores, in terms of number and percent correct, for each section of the Pre-CPT for the total group of field test examinees, the total target group, and the target group excluding examinees who indicated that they spoke Chinese at home.

Table 5
Field Test Means for Total Group and Sub-Groups

Section	Total Group[1]	Target Group[2]	Target Sub-Group[3]
Listening			
# Correct	52.59	49.73	44.93
(Std. Dev.)	(14.20)	(13.12)	(9.45)
% Correct	69%	65%	59%
# Examinees	266	120	91
Reading			
# Correct	43.72	40.92	39.43
(Std. Dev.)	(12.67)	(11.59)	(10.73)
% Correct	69%	65%	63%
# Examinees	254	114	86
Structure			
# Correct	17.35	14.91	13.70
(Std. Dev.)	(7.43)	(6.45)	(5.54)
% Correct	58%	50%	46%
# Examinees	262	118	89

[1]The Total Group includes all examinees who participated in the Pre-CPT field test.
[2]The Target Group includes all examinees who indicated that they were in either the 3rd or 4th year of high school or had earned between 4 and 9 college credits in Chinese.
[3]The Target Subgroup includes only those examinees of the Target Group who indicated that they did NOT speak Chinese at home.

To determine if an item was too easy or too difficult to include on the test on the basis of the p-values obtained for the entire group, it was necessary to remember: 1) that the optimal mean p-value for a multiple choice test with 4-choice items is 62.5% or somewhat higher (Crocker & Algina, 1986, p. 313), 2) that the field test version of the Pre-CPT was given slightly past mid-year while most examinees would take the test towards the end of the school year when their abilities should be greater, and 3) that the ability level of the entire group was above the ability level of the target group. Thus, the range of acceptable p-values for the final form needed to be modified.

To adjust the range of p-values derived for the total group in order for it to be appropriate for the target group, the difference between the mean performance in terms of percent correct on each section for the total group and the target subgroup without Chinese speakers (given in Table 5) was calculated. This difference was then added to 62.5, the appropriate lower bound mean p-value for a multiple-choice test. For Listening Comprehension, the difference was 10; thus, the optimal p-value based on the results of the entire group became 72.5. For Reading Comprehension, the difference was 6; thus, the optimal p-value became 68.5. For Structure, the

difference was 12, and the optimal p-value became 74.5. Items in each section with total group p-values within 20 points above or below these means (72.5 for Listening Comprehension, 68.5 for Reading Comprehension, and 74.5 for Structure) were thus considered appropriate in terms of difficulty to be retained on the final form of the Pre-CPT[1].

There were no problems meeting these selection criteria for the Listening and Reading Comprehension sections of the test. There were enough items on the field test that required no revisions (on the basis of the analysis of discrimination and functioning of the distractors) to select the required number for the final form. However, more than half of the Structure items were too difficult. To reduce the difficulty level of these items, several changes were made. First, one passage containing four difficult items (of five items total) was deleted. Second, modifications were made to the remainder of the problematic items, without altering the original passages. These modifications included revising distractors which appeared too unfamiliar or too attractive. In some cases, the difficult items were simply replaced by newly devised items. In this process, the original missing word was re-inserted, and another word was omitted in its place to create a new item thought to be more appropriate to the target group's level of ability.

4.3 Qualitative Input

Input from test supervisors was both helpful and encouraging. Many teachers felt that the Listening Comprehension section would be facilitated by a kind of non-graded introductory lead to familiarize the examinees with the speakers' voices. In response to this suggestion, the final version of the taped directions to the Listening Comprehension section includes an introduction in which three of the Chinese speakers read a sample monologue aloud to introduce students to their voices.

Supervisors also commented on the interval in which examinees had to answer the questions in the Listening Comprehension section. On the field test form, there were 12 seconds for examinees to respond to each item. Supervisors pointed out that items containing lengthy options might require extra time for examinees to read them. Accordingly, in the final version, three extra seconds were added to this pause time for items with longer options.

In the field test, the Structure section preceded the Reading Comprehension section and was included as a part of the Reading Comprehension section (as in the original CPT). Some test supervisors suggested that it be placed after the Reading Comprehension section and be considered a separate section of the test. This suggestion was incorporated into the final version of the test.

---

[1] The Rasch analysis program which was used later for test equating was not available at this stage in the test development process. Had it been used, decisions regarding selection of appropriate items in terms of difficulty would have followed a rather different procedure based on the calibrated person ability and item difficulty measures, and on item fit statistics.

Test supervisors made a number of comments about the Reading Comprehension section. One suggestion was to supply more language materials from outside mainland China for a better balance of selections. Some supervisors provided corrections to the Chinese and English texts. They pointed out certain grammatical structures which sounded too colloquial and required revision to fit the written style of Chinese; certain versions of Chinese characters that were not in the correct traditional or simplified forms; and some inconsistent uses of the Chinese Pin-yin system for transcribing proper names in the English text.

4.4 Revision of the Test

On the basis of both the quantitative item analysis and comments from test supervisors, revisions were made to the field test form of the Pre-CPT. In addition to those mentioned in the preceding paragraphs, the following changes were also made.

In the Listening and Reading Comprehension sections, the few poorly performing items were deleted, as well as items that were either too difficult or too easy for the target group of examinees. In a very few cases, minor revisions were made to certain items that showed a poor performance due to an inappropriate distractor or a specific lexical or syntactic item in the stimulus. The target number of items, 55 for Listening Comprehension and 45 for Reading Comprehension, were achieved in this way.

In the Structure section, after the revisions based on the statistical analyses, the total number of items was reduced from 30 to 25 by deleting a whole passage with five test items.

There was one final major revision: the Structure section became entirely independent of the Reading Comprehension section and was placed at the end of the test. This was done for a number of reasons. First, it was suggested by some of the test supervisors. Second, other test supervisors expressed the opinion that they would like to have an option not to give the Structure section to some groups of students. Third, placement at the end of the test appeared to be appropriate since this section was the most difficult.

5. The Norming Administration of the Pre-CPT

After revisions to the Pre-CPT following field testing, the test was administered in Chinese language programs throughout the country for norming purposes between late April and early June, 1991. This chapter describes the rationale behind, procedures for and results of the norming administration.

5.1 Rationale for the Norming Administration

The norming administration had two objectives. The first was to provide preliminary national norms to be used in interpreting test scores. The second was to equate the Pre-CPT with the CPT so that scores on both tests could be interpreted on a common scale, to be called the CPT Scale.

To meet the first objective, it was necessary to invite as large a group as possible of examinees typical of the target population of the Pre-CPT to participate in the norming administration. Invitations to participate were sent to over 40 college and university programs and 80 high school programs. Through a presentation on the project by CAL staff at a conference sponsored by the Eastern Association of Chinese Schools, some weekend Chinese schools also participated. **No students that had participated in the field testing were allowed to participate in the norming administration.** The list of schools that participated and the number of students from each is found in Appendix C: Pre-CPT Norming Administration Participants.

To meet the second objective, it was necessary to have common items on both the Pre-CPT and the CPT. These common items would serve as anchor items to link the two tests. Using an analysis of the performance of 174 beginning level examinees (i.e., examinees who indicated that they were in a first year college level Chinese language course) on the current CPT, items that were not too difficult for this group yet still discriminated fairly well for all CPT examinees were chosen for inclusion as equating items on the Pre-CPT norming administration form. Ten items for Listening Comprehension and 10 for Reading Comprehension were selected.

For the Structure section, selecting anchor items was not as straightforward. The CPT has 35 structure items utilizing two separate item types. The first asks examinees to indicate at what point within a Chinese sentence a certain Chinese character would be correctly placed. Four possible locations are indicated. The second item type is a single-sentence multiple-choice cloze that asks examinees to complete the missing portion of a sentence with one of four options. However, in the Pre-CPT, only one item type, a standard multiple-choice cloze, is used in the Structure section. Here, examinees are presented with a paragraph with five words missing. For each missing word examinees are asked to choose the best completion from among four options. The items on the CPT most similar to these were the 20 single-sentence cloze items. Unfortunately, most of these were very difficult for both the beginning-level CPT examinees and the entire CPT population. Of the 20 single sentence items, only six appeared potentially appropriate for the Pre-CPT target group population. Thus, only these six items could be used

to link the Structure sections of the two tests. On the norming version of the Pre-CPT, these six items were separately presented to examinees as the first part of the structure section of the test. The second part contained the 25 paragraph-level items developed for the Pre-CPT.

Table 6 shows the format for the norming administration version of the Pre-CPT.

Table 6
Organization of Norming Administration Version of the Pre-CPT

Section	Total Time	Number of Items	Part	Format of Stimulus	Number of Items
Listening Comprehension	30 min (approx)	65			
			One	Utterances	22
			Two	Dialogues	33
			Three	Monologues	10
Reading Comprehension	55 min	60			
				Nonlinear Text	12
				Signs	10
				Passages	33
Structure	25 min	31			
	(5 min)			Single-Sentence	6
	(20 min)			Paragraphs	25

5.2 The Results of the Norming Administration

651 examinees participated in the norming administration. Because the Pre-CPT is a general proficiency test, no effort was made to exclude examinees studying Chinese who spoke Mandarin or another Chinese language at home. A background questionnaire (Appendix B) was completed by the vast majority of examinees. It revealed that 48.8% were male while 51.2% were female (with nine giving no response). Students in college comprised 47.5% of the total population, followed by students in public high school (36.3%), students in private high school (13.1%), and students enrolled in other schools; i.e., weekend Chinese language schools (3.2%).

Of those responding to the question about ethnicity (96.2% of the total), 69.2% stated they were Chinese. We believe that this proportion of ethnic Chinese is fairly typical of the combined advanced level high school and first year college test population. Although there are regional differences in the proportion of ethnic Chinese studying the Chinese language throughout the United States, it is reasonable to assume that overall a majority of high school students taking Chinese are of Chinese ethnic background, given the difficulty of Chinese relative to Spanish or French for the American student population. For the same reason, many first year students of Chinese in college are probably ethnic Chinese also. Of the 95.5% who responded to the question about their home language, 14.8% indicated that they spoke Mandarin Chinese, 32.2% indicated Chinese but not Mandarin, while a slight majority (53.1%) indicated they did not speak

any Chinese at home. Table 7 presents a summary of the demographic data for the total norming sample.

Table 7
Demographic Information on
Participants in the Norming Administration

Total Number	651
Ethnicity	
Chinese	69.2%
Non-Chinese	30.8%
Languages	
Mandarin spoken at home	14.8%
Other Chinese language spoken at home	32.2%
No Chinese spoken at home	53.1%
Level of Chinese Instruction	
High school	52.5%
College	47.5%

It may be noted that there were some differences in this sample between the ethnic composition of students studying Chinese at high schools and at colleges. 85.4% of the public high school students indicated that they were of Chinese ethnicity, and 66.2% of them spoke Chinese at home (16% Mandarin and 50.2% Chinese, but not Mandarin). Of the private high school students studying Chinese, the situation was the opposite. Only 23.5% were of Chinese ethnicity with only 16.3% speaking Chinese at home (3.8% Mandarin; 12.5% Chinese, but not Mandarin). The college students who took the Pre-CPT during the norming administration were also predominately Chinese (68.3%), but unlike the public high school students, less than half (40.9%) spoke Chinese at home (14.5% Mandarin, 26.4% Chinese but not Mandarin).

It is not surprising to find that the majority of public school students studying Chinese as an elective are Chinese and speak Chinese at home. Nor is it surprising to find beginning level college students with the distribution described above. Thus, given these figures, CAL staff became concerned about how best to construct norm tables. Separate norms for high school and college students would be helpful to test users, but would separate norm tables for Chinese and non-Chinese speakers be helpful? After further analysis (detailed in Section 7.2 of this report), it was decided that two separate Pre-CPT norm tables would be provided. One would be for all examinees, and the other for non-Chinese-speaking students only, since the Pre-CPT appears most suitable for English-speakers, and we feel the greatest number of students taking the Pre-CPT in the future will be English-speaking (the CPT appears to be the more appropriate test for Chinese-speaking students at all levels). The final norm tables appear in Appendix D. CAL staff has also provided, as an aid to test users, mean scores and standard deviations for all separate subgroups in the norming administration. (These are given in Table 13, Section 7.2 of this report.)

Results of a classical item analysis are given in Table 8 below, which presents the summary of descriptive statistics by section.

Table 8
Statistics from the Norm Test Administration
Descriptive Statistics

Section	Number of Examinees	Number of Items	Mean Score	Std. Dev.	Reliability	Mean P-value	Std.Dev. P-value
Listening	648	65	52.41	11.12	.94	.81	.11
Reading	642	60	44.51	10.46	.92	.74	.13
Structure	635	31	21.06	6.87	.89	.68	.12

These results show that the subtest reliabilities were very high. While the mean p-value for the structure and reading items was appropriate for a multiple choice test, the Listening Comprehension section may have been somewhat easy for this group. This was probably due to the large number of native speakers of Chinese in the sample. These native speakers would not have enjoyed real advantages in the Reading and Structure sections, which require the examinee to be able to read Chinese characters. However, they would have a decided advantage on the Listening section, especially if they spoke Mandarin at home.

6. Building a Common Scale for the Pre-CPT and the CPT (The CPT Scale)

Besides providing norming information, a second goal of the norming administration was to build a common score scale for the CPT and the Pre-CPT based on the administration of common items in each section of the test (i.e., Listening, Reading and Structure). In psychometric literature, joining two tests which measure the same construct but are targeted at different ability levels is called "vertical equating." CAL staff first surveyed the current literature to determine what approaches would be most appropriate for the Pre-CPT/CPT situation. Although the literature revealed a number of different methods to equate the two tests, ultimately the Rasch model was chosen to accomplish the equating.

The Rasch model is a probabilistic measurement model which can be classified in the family of models based on latent trait theory (item response theory or IRT). IRT is a modern approach to measurement which seeks to overcome the limitations of classical measurement theory. In classical theory, one serious limitation is that item characteristics (particularly the difficulty of the item) are dependent on the ability of the group of examinees to whom the item was administered. The same item could be labeled difficult when administered to group A, but easy when administered to group B. The measurement of ability (an examinee's score) is likewise dependent on the specific test (group of items) the examinee took. The same examinee can appear strong on Test A but very weak on Test B, depending on the overall difficulty of the two tests. Another drawback to the classical approach is that there is no way to relate measures of examinee ability and measures of item difficulty on the same scale.

IRT models overcome these limitations. They place measures of the difficulty of an item (item difficulty) and the ability of a person (person ability) on the same scale. They also allow for the measurement of item difficulty indices independent of the ability of the sample of examinees who take the test, and for the measurement of an examinee's ability independent of the test items administered (Hambleton & Swaminathan, 1985, p. 11)[2].

The Rasch model was chosen as the tool with which to equate the two tests for several reasons. First, the Rasch model has been widely used in language testing. Because of its relative simplicity, it is the most-widely applied of all IRT models. Second, it is often considered the most appropriate model when small numbers of examinees are available for estimating person ability and item difficulty (as is the case in testing the uncommonly-taught languages). It is expected that less than 1000 examinees will be taking the Pre-CPT annually. Finally, a flexible, user-friendly, high capacity software program for Rasch measurement (BIGSTEPS) (Wright & Linacre, 1991b) recently became available for personal computers. CAL staff used BIGSTEPS

---

[2] For more on Item Response Theory, see Hambleton & Swaminathan (1985) and Hambleton et al. (1991). For a good introduction to latent trait theory in the language testing context, see Stansfield (1985) and Henning (1987).

to conduct the research presented in this report. BIGSTEPS is also used in the operational program to score the tests.

6.1 Concurrent Calibration of the Pre-CPT and the CPT

In the early applied IRT methodology, because of the limitations of available computer programs, when two tests were to be equated, the two tests would be separately calibrated. In other words, item difficulty measures and person ability measures would first be estimated for Test A, and then for Test B. Because the origin of the two scales would be different, the measures for one test would have to be converted into the metric of the other.

BIGSTEPS now permits the concurrent calibration of tests being equated. This means that only one score metric is produced, which is the same for all the tests being equated. Concurrent calibration was made possible because BIGSTEPS allows items for which the examinee gives no answer to be treated as "unreached" rather than incorrect. In this application, the two data sets containing the responses to both the CPT and the Pre-CPT were combined. Responses to the common items formed a single column in the combined data set. For the rest of the columns, items unique to the CPT contained blanks for the Pre-CPT examinees, and items unique to the Pre-CPT contained blanks for the CPT examinees. All blanks were treated as "unreached" rather than as incorrect responses.

To prepare for the equating, the CPT data bank was first updated to include all examinees who had taken the test as of June, 1991. Since some examinees take the CPT more than one time, the database used for the analysis was a subset of the complete database, in order that each examinee would appear in the calibration sample only once. Also, examinees who failed to take one of the CPT subtests were excluded from the database submitted for analysis of that subtest. Thus, for the concurrent calibration, the following numbers of CPT examinees were used: 1697 for Reading, 1736 for Structure, and 1697 for Listening. All 651 examinees who took the norming administration form of the Pre-CPT were included in the analysis.

The first step in determining whether vertical equation of the two tests is appropriate is to examine the extent of the relationship between the item difficulty calibrations when the common items are calibrated separately. In item response theory, measures of item difficulty are independent of the group of examinees used to calibrate the items. Thus, the common items should receive the same difficulty measurements (within statistical error and on scales centered at different points) whether calibrated with the Pre-CPT norming sample or the CPT sample. The two calibrations should be highly correlated. Table 9 shows the correlations between the common items separately calibrated. The correlations have been disattenuated to account for errors of measurement.

Table 9
Correlations of the Difficulty Values of the Common Items Calibrated Separated for the Pre-CPT and the CPT Populations

Section	Number of Common Items	Correlation (Disattenuated)
Listening	10	.92
Reading	10	.92
Structure	4	.94

Table 9 indicates that only four of the six items originally selected from the CPT to serve as anchor items for the norming administration of the Pre-CPT were actually appropriate for the Pre-CPT sample. The two deleted items proved much too difficult for the Pre-CPT sample of students. They did not differentiate between performance levels in the Pre-CPT sample of students as they could in the CPT sample. Thus, it was inappropriate to use them in equating the Structure section of the two tests. (It may be noted that the Structure section is optional for the Pre-CPT and that all CPT Structure items have been removed from the final form of the Pre-CPT.)

We also examined whether there would be differences in model fit under separate versus concurrent calibration. Model fit is an important factor in appropriate Rasch model use. The probabilistic Rasch model posits that examinees have a fifty percent chance of getting an item correct when the item's difficulty is the same as the examinee's ability. When this is not the case (for example, when a low ability examinee gets a difficult item correct, or when item difficulty and student ability are close and the examinee gets the item incorrect), there is misfit. In the BIGSTEPS program, misfit is indicated through the calculation of four fit indices. Two of the indices are for OUTFIT. These indices are heavily influenced by unexpected responses by persons on items far from the person's ability level. Two of the indices are for INFIT. These indices are weighted in such a way that they are less influenced by unexpected behavior on items far from the person's ability level, and are thus more sensitive to unexpected behavior affecting responses to items near the person's ability level. These indices may be either positive or negative. Positive misfit indices for INFIT indicate "noise" in the data; the larger the amount, the more instances (and/or the greater severity) of examinees not performing as expected at items near their ability level (e.g., missing items that they should have gotten correct). Positive misfit indices for OUTFIT indicate the presence of unexpected outliers; the larger the amount, the more instances (and/or the greater severity) of examinees not performing as expected on items far from their ability level (e.g., getting items correct which are much greater in difficulty than their ability level or getting items incorrect which should be very easy for them). Negative values for both misfit statistics indicate unusually predictable responses to an item. In other words, performance on these items tends be very consistent and the items can be seen as providing redundant measurement information. The extreme case of negative misfit would occur if the exact same item appeared twice in the test.

There is no straightforward methodology for the analysis and interpretation of misfit. Each situation must be looked at individually. In our case, we had certain constraints. First, we could not change or discard any of the items on the CPT. All of them needed to be used since the CPT was a pre-existing test. Second, although it is often done, we did not feel comfortable discarding any misfitting examinees from either the CPT or the Pre-CPT database to increase overall model fit. There were already well over 2000 examinees who had taken the CPT in its operational program, and we believe both them and those who took the Pre-CPT during the norming administration to be representative of all the types of examinees in the two operational testing programs. Given the large and disparate sample sizes for the CPT and the Pre-CPT populations, the criterion used was the infit and outfit mean square statistic provided in the BIGSTEPS output, which is not sensitive to sample size. Following standard practice, an item was considered misfitting if both of the mean square fit statistics were greater than 1.20 (positive) or less than .80 (negative). There are three sets of items to be considered: those unique to the Pre-CPT, those unique to the CPT, and the common anchor items. Table 10 indicates the number of the anchor items that were misfitting under the separate Pre-CPT and CPT calibrations, and misfitting under the concurrent calibration.

Table 10
Number of Misfitting Anchor Items Under Separate and Concurrent Calibrations

	Pre-CPT Separate Calibration	CPT Separate Calibration	Concurrent Calibration
Listening (10 items)			
>1.20	1	0	0
<.80	1	0	1
Reading (10 items)			
>1.20	1	0	0
<.80	0	0	0
Structure (6 items)			
>1.20	1	0	0
<.80	0	0	0

Table 10 indicates that fit was not a problem for the anchor items. None of the anchor items (all of which came from the CPT) in any section were misfitting on the CPT, though two of the anchor items were misfitting in the Listening section of the Pre-CPT, and one was misfitting in the Reading and Structure sections. However, when these anchor items were concurrently calibrated using the entire sample, only one (in the Listening section) remained misfitting.

Table 11 shows the number of the items unique to each test misfitting under separate and concurrent calibration.

Table 11
Number of Misfitting Items Under Separate and Concurrent Calibrations

	Separate Calibration	Concurrent Calibration
Unique Pre-CPT Items		
List (55 items)		
>1.20	4	4
<.80	2	2
Read (50 items)		
>1.20	2	2
<.80	0	0
Str (25 items)		
>1.20	2	2
<.80	0	0
Unique CPT Items		
List (50 items)		
>1.20	4	4
<.80	3	3
Read (45 items)		
>1.20	4	4
<.80	1	1
Str (29 items)		
>1.20	3	3
<.80	0	0

Table 11 indicates that the number of misfitting items under separate and concurrent calibration was exactly the same. Upon closer analysis, all of the items misfitting under each calibration were exactly the same, and if their mean square INFIT and OUTFIT statistics differed at all, it was by a maximum of only .01 logits. Tables 10 and 11 indicate that concurrent calibration effects only the common items and not the unique items on the tests to be equated.

6.2 Building the CPT Scale

As a result of the concurrent calibration, all examinees now had an ability score on the same scale. This ability score is no longer a "number right" score, but an estimate of the person's ability along the continuum of the latent trait (the construct being measured) reported in terms of logits (a ratio in natural log odd units--see Wright and Stone, 1979) centered at 0 (the average item difficulty) and extending from about -6 to +6[3]. An examinee's ability score in logits is

[3] Technically, ability cannot be estimated for examinees who get all items either correct or incorrect. For CAL's Chinese language testing program, BIGSTEP's default estimation

defined as the point on the item difficulty scale where the examinee has a fifty percent chance of getting the answer correct. Thus, an examinee with an ability of 1.00 logits has a fifty percent chance of getting an item with a difficulty level of 1.00 logits correct. This examinee's chances of getting an item at 0.00 logit is greater than 50%, while for an item at 2.00 logits it is less.

This logit scale, however, can be changed by any linear transformation without losing its linear quality. Since CAL's Chinese testing program has traditionally been interpreted in terms of norms, CAL staff decided that the scaled score for the CPT and Pre-CPT should reflect a norm-referenced interpretation. The scale point of 100 was chosen to reflect the mean of the new CPT Scale, with one standard deviation to be equal to 20 points. 100 would then be interpreted as the average ability score for the both Pre-CPT and CPT examinees; that is, for all students participating in CAL's Chinese language testing program. An examinee receiving a score of 120 would be one standard deviation above this mean; an examinee receiving a score of 80 would be one standard deviation below this mean.

Below are the three equations used to transform ability estimates in logits to scaled scores. Scaled scores are rounded to the nearest integer.

Listening Comprehension
CPT Scale Score = 78.76 + (17.70xLogit Score)

Reading Comprehension
CPT Scale Score = 81.93 + (17.54xLogit Score)

Structure
CPT Scale Score = 91.78 + (18.69xLogit Score)

6.3 Building the Norming Tables

To help test users interpret the meaning of scores on the Pre-CPT and CPT, CAL developed norm tables based on the scaled scores. For Pre-CPT users, there are two norm tables. Both were divided based on High School students and College Students (1st Year of Study). The first reflects the performance of all examinees participating in the norming study. The second is based only on the performance of the examinees who indicated that they did not speak Chinese at home. For the CPT, norms have always been divided according to the college level course designations "Beginning," "Intermediate" and "Advanced." These designations are derived from self-reported information provided by examinees and refer primarily to the course the examinee is enrolled in (most typically completing) at the time the test is taken. For those not enrolled in

---

procedure was used for extreme scores. Although this occurred very infrequently on the CPT, it was rather frequent for native speakers on the Pre-CPT, occurring for 9.4% of the examinees for Listening, 3.4% for Reading and 6.9% for Structure.

any Chinese language class at the time of the test, it refers to the level of the last Chinese language course completed.

Before the norm table for the Pre-CPT could be built, the items that would appear on the final form of the test needed to be selected. This procedure is discussed in Chapter 7. The complete norm tables are presented in Appendix D.

7. The Final Form of the Pre-CPT

7.1 The Selection of Items

As mentioned in Chapter 5, there were more items on the Pre-CPT than had been envisioned for the final form of the test. The 65-item Listening Comprehension section contained 10 common items, of which only five could remain. The performance of the ten items was inspected, and one item with low discrimination was removed. The remaining nine common items were psychometrically acceptable. Five of these, which represented various levels of difficulty, were chosen to remain in the final form.

To shorten the test, it was decided that the final number of listening items would be 50. Thus, 10 more items were to be deleted. None could be from Part Three (monologues), which had only 10 items. Thus, although all the remaining items were technically good, 10 items that were either relatively easy, duplicated test content or had a relatively lower discrimination value were eliminated. This left 50 listening items: 20 in Part One, 20 in Part Two and 10 in Part Three. Five of the 50 items serve as anchor items that are common to both the CPT and Pre-CPT.

Five of the ten Pre-CPT Reading Comprehension items that also appeared on the CPT needed to be deleted. All were technically sound, but only a sample of those representing both signs and passages at various levels of difficulty could be kept. In order to make a total of 50 items in this section, five additional items were deleted. At this point, it was decided that all the nonlinear text and sign items should be kept. Therefore these five additional items came from the passages and were deleted on the basis of being less authentic (in that the text had been quite modified to be made appropriate to the level of the Pre-CPT) or having relatively lower discrimination indices in the traditional item analysis that was run. In its final form, the 50 reading items include: 12 nonlinear text items, 8 signs and 30 passages of various lengths. Among the 50 items, 5 are anchor items common to both the CPT and Pre-CPT.

To link the Structure section of the two tests, 6 single-sentence cloze type items had been taken from the CPT and placed in the first part of the Pre-CPT Structure section (see Chapter 5). In revising the Pre-CPT Structure section, it was decided to delete these single-sentence cloze items and to keep the paragraph cloze part intact. This makes the Pre-CPT Structure section easier to administer and interpret[4]. However, there are now no items common to both tests in the Structure section.

CPT Scale scores for norming purposes were estimated for the Pre-CPT examinees by using only the items that were included on the final form of the Pre-CPT. To do so, all these items

---

[4] It is easier to administer because with only a single item type, only one set of directions and sample items is needed. This also makes the test shorter, which means it takes less time to administer. It is easier to interpret because it is easier to understand the meaning of a score based on one item type than a score based on two different item types.

were viewed as "anchor" items by the BIGSTEPS program. In other words, the program did not estimate item calibrations, but used the calibrations stemming from the concurrent calibration. The estimates of person ability in logits were then converted to the CPT Scale score using the formulae given in Chapter 6.

Scale scores for the CPT were likewise determined by a separate calibration of all examinees in the CPT database (over 2200), not just those used in the concurrent calibration. In other words, if an examinee took the CPT more than once, he or she received an ability estimate for each occasion the test was taken.

A misfit analysis was conducted on those items remaining on the Pre-CPT based on the scoring calibration using the same criterion as used previously. An item was marked as misfitting if both its mean square INFIT and OUTFIT fit statistics were greater than 1.20 or less than .80. Table 12 shows the number and percent of misfitting items. For purposes of completeness, similar data is presented for the CPT, based on its scoring calibration.

Table 12
Number and Percent of Misfitting Items on the Final Forms of the Pre-CPT and CPT

	Pre-CPT	CPT
Listening Comprehension		
>1.20	1	4
<.80	3	1
TOTAL	4/50 8%	5/60 8%
Reading Comprehension		
>1.20	2	3
<.80	2	1
TOTAL	4/50 8%	4/55 7%
Structure		
>1.20	2	1
<.80	0	1
TOTAL	2/25 8%	2/35 6%

Some misfit is expected. In the Rasch model, a measure is generally regarded as appropriate (i.e., that the items all measuring the same underlying trait), if less than 10% of the items are misfitting. Thus, Table 12 indicates that on both tests and in all sections, the items conform to the underlying variable in each trait.

7.2 The Final Norming Tables

The final norming tables are presented in Appendix D: Norming Tables for the Pre-CPT and CPT. In addition to these tables, the *CPT/Pre-CPT Combined Test Interpretation Manual* also presents the means for the various subgroups in this norming sample. These are printed in Table

13 on page 36. For the Pre-CPT, norming subgroups are divided by high school and college course level and whether the examinees speak Mandarin, another Chinese language, or English at home. For the CPT, norming subgroups are divided by college course level and whether Chinese or English is indicated as the native language.[5] The college course level is self-reported. At the time examinees take the CPT, they indicate what is the highest level of Chinese language course (not literature) they are presently enrolled in. Generally students completing a first year course indicate "Beginning," students completing a second year course indicate "Intermediate," and students completing a third year (or higher) level course indicate "Advanced." Note, however, that not all examinees take the CPT at the end of an academic year.

In the table, the means for each section are given on the first line in bold. Underneath each mean is its standard deviation. On the bottom line, in parentheses, is the number of examinees in the subgroup. For the Pre-CPT, means for subgroups with less than 10 members were not calculated.

The means in Table 13 indicate that, for the norming population, the two tests appear to give appropriate results. Thus, we see that within any level there is a wide divergence in performance on the CPT between Chinese and English speakers, and, for the Pre-CPT, between Mandarin and non-Mandarin speakers of Chinese as well. The means also reveal that for English speakers, the two tests show consistent progress as levels increase, and that mean scores on the Pre-CPT for second semester college students were very close to the mean scores on the CPT for the beginning level students, which would be expected. Also, as may be expected, fourth (and especially third) year high school students do not do quite as well as second semester college students. The only unexpected result occurs in the means of the Listening Comprehension section for English speakers between the third and fourth year of high school. The mean for the third year is 83.59. The mean for the fourth year would be expected to be higher, but it is slightly lower (81.22). Perhaps also unexpected are the high scores for the fourth semester college students on the Pre-CPT, which exceed the Intermediate level scores on the CPT. It must be remembered, however, that the Pre-CPT means are based on a very small number compared to the CPT means, and that all the Pre-CPT examinees took the test at the end of the year, whereas some Intermediate CPT examinees may have taken the test at other times during the school year. For example, students taking the CPT for entrance into a study abroad program typically take the CPT in the early spring.

---

[5] The CPT answer sheet, used since the beginning of the CPT program, does not capture information on the type of Chinese spoken by the examinee.

Table 13
Means, Standard Deviations and Number of Examinees
by Level and Language Background
for the Pre-CPT and the CPT
(in CPT Scale Scores)

PRE-CPT MEANS TABLE

LEVEL	LISTENING			READING			STRUCTURE		
	Mandarin	OthrChin	English	Mandarin	OthrChin	English	Mandarin	OthrChin	English
HIGH SCHOOL									
	--	97.00	72.29	--	97.64	75.97	--	96.55	76.91
2nd Year	--	27.68	19.45	--	27.68	15.74	--	29.52	21.22
	--	(11)	(34)	--	(11)	(31)	--	(11)	(32)
	129.22	106.69	83.59	115.15	103.79	80.49	124.00	108.96	87.43
3rd Year	18.60	24.16	20.95	19.37	23.20	18.32	16.68	23.22	22.64
	(27)	(72)	(69)	(27)	(72)	(68)	(27)	(72)	(69)
	139.80	123.79	81.22	118.30	121.56	86.94	119.00	130.00	92.03
4th Year	20.21	17.69	28.65	39.69	17.84	21.49	25.75	20.25	26.03
	(10)	(34)	(36)	(10)	(34)	(36)	(10)	(34)	(36)
COLLEGE									
	133.90	107.19	88.25	105.32	96.05	90.37	114.33	104.40	92.93
2nd Sem	18.70	21.63	21.30	21.04	22.20	19.48	22.96	26.96	21.73
	(41)	(75)	(147)	(41)	(77)	(148)	(40)	(75)	(140)
	--	--	103.46	--	--	108.83	--	--	112.58
4th Sem	--	--	19.90	--	--	24.76	--	--	26.76
	--	--	(24)	--	--	(24)	--	--	(24)

CPT MEANS TABLE

LEVEL	LISTENING		READING		STRUCTURE	
	Chinese	English	Chinese	English	Chinese	English
	119.18	92.11	102.18	89.00	108.82	89.83
Beginning	20.00	15.47	18.78	14.08	27.34	12.77
	(17)	(389)	(17)	(363)	(17)	(391)
	111.21	98.12	105.36	98.56	111.97	97.48
Intermediate	18.52	15.04	19.29	15.35	21.73	15.34
	(39)	(934)	(39)	(932)	(39)	(935)
	133.00	111.10	134.00	115.11	146.27	111.12
Advanced	16.89	16.41	20.48	18.62	32.08	18.65
	(26)	(634)	(26)	(639)	(26)	(639)

For native Chinese speakers, the Pre-CPT also appears to consistently show expected differences in levels of Chinese language instruction, with the exception of the Structure section scores of Mandarin speakers. (It should be remembered, however, that the mean of the fourth year students is based on only 10 examinees in that subgroup.) The Pre-CPT means for native Chinese speakers from the norming sample may also appear surprising when comparing native speaking students in the third and fourth year of high school with native speaking second semester college students on the Pre-CPT, or Chinese speaking examinees at the Beginning and Intermediate levels on the CPT. However, this may also be due to the fact noted above, that

there were many native speaking examinees with perfect scores, particularly for listening comprehension, whose scale score estimates were produced by BIGSTEPS, though a genuine ability estimate for such examinees is not possible to compute. This has no doubt inflated the means for the native-speaking Pre-CPT examinees. Another factor may be that there are true differences in linguistic ability between native speaking students who study Chinese in high school and those who study the language in college. Perhaps native speakers who choose to study Chinese in high school tend to be stronger in the language than those in college. Home support for the language may be stronger for high school students than for college students living away from home in an English-speaking environment. Finally, for the CPT, we don't know how many of the sample spoke Mandarin or another Chinese language. Note that the data above is only intended to describe the norming population for the tests.

7.3 Difficulty of the Test

Table 14 gives the mean item difficulty for each test and each section in terms of the CPT Scale score.

Table 14
Means (and Standard Deviations)
of Item Difficulties
in CPT Scale Scores

Test	List	Read	Str
Pre-CPT	60.53 (13.98)	65.44 (14.73)	74.62 (11.53)
CPT	98.94 (22.13)	97.72 (17.54)	103.58 (18.74)

One way to interpret these mean item difficulties is to say that a person with the corresponding ability level taking each section of the test would get 50% of the items correct. Since in general examinees feel more comfortable with tests on which they can answer more than half of the items correctly (60% is generally seen as failing in standard classroom exams), these means should be viewed as limits on appropriate examinee ability to take each test. Thus, examinees at an ability level in the 80's would find the CPT very difficult while examinees at an ability level over 100 would find the Pre-CPT very easy.

Given Tables 13 and 14, the Pre-CPT appears to be the more appropriate test for English speaking students at all high school levels and at the beginning level of college instruction. These students will find the CPT too difficult. On the other hand, for all native speaking Chinese students (except perhaps speakers of Chinese languages other than Mandarin in the second year of high school Chinese or a beginning level college course), the CPT would be sufficiently challenging and psychometrically appropriate.

Table 13 clearly shows that the Pre-CPT was easier for Chinese speaking students than for English speaking students. However, it is not clear that the items required the same kinds of skills for the two groups. One way to examine this using the Rasch model is to compare the item difficulties when calibrated separately for the two different groups. Although the absolute difficulty values will be different, there should be a high correlation between the two.

On the complete norming version of the Pre-CPT, the correlation (disattenuated to account for errors of measurement) between the item difficulties when separately calibrated on 327 non-Chinese speakers and on 292 Chinese speakers for listening comprehension was .87 (65 items); for reading comprehension, .95 (60 items), and for structure, .86 (31 items). These figures provide evidence that the items were functioning similarly for each group.

8. Psychometric Properties of the Pre-CPT and CPT

8.1 Reliability

The reliability of a test is the extent to which it yields consistent results. Thus, high test reliability is desirable. A test may, however, have different reliabilities in different populations. Since a large number of Chinese-speaking students participated in the norming administration of the Pre-CPT, we can examine the reliability of the test for both Chinese-speaking and non-Chinese speaking examinees.

Table 15 gives the Kuder-Richardson (Formula 20) reliabilities of the section scores from the Pre-CPT based on the total norming population, and on two subpopulations. Reliabilities are calculated from the norming administration data using only those items remaining on the final form. Table 15 also gives the reliabilities for the CPT, which are those published in the original *CPT Test Manual* (Wang & Stansfield, 1988). These are based on 479 examinees who took the CPT between 1984 and 1987[6].

Table 15
Reliability of the
Pre-CPT and CPT by Section and Group

Section	Total Group	Chinese Speakers	Non-Chinese Speakers
Pre-CPT			
Listening Comp	.94	.94	.92
Reading Comp	.93	.92	.93
Structure	.88	.90	.90
CPT			
Listening Comp	.89		
Reading Comp	.93		
Structure	.83		

It may be noted that the reliability of the Pre-CPT remains high across subsections and subgroups, even though the number of items in each section was reduced in preparing the final form. These reliability coefficients, based on a subset of items given in the norming administration, should be updated once there is a large number of examinees in the database of the operational program. The figures presented in Table 15 indicate that both the Pre-CPT and

[6] The reliabilities calculated for the CPT by the BIGSTEPS program based on the complete updated database, are comparable: Listening (.90), Reading (.90) and Structure (.83). No division was made between the Chinese speakers and the Non-Chinese speakers in the original *CPT Test Manual*.

the CPT are highly reliable tests which can be used with confidence by programs needing trustworthy measures of Chinese language proficiency.

8.2 Precision of Measurement

Any measurement of an individual's ability involves a degree of error. The smaller the error, the higher the precision of the measurement. In Classical Test Theory, the degree of error is usually indicated by the standard error of measurement (SEM). One of the limitations of this approach, in which the SEM is calculated using the test's reliability and the sample test score variance, is the assumption that the SEM is the same for all examinees. It is well-known, however, that test scores are unequally precise measures for examinees at different levels of ability (Hambleton *et al.*, 1991). Item Response Theory brings this question of the precision of the examinee's ability estimate to the forefront. IRT estimates of precision for each examinee's score are a function of 1) how many items an examinee attempts, and 2) how far the difficulty level of the items are from the examinee's level of ability. Optimum precision occurs when an examinee attempts a sufficient amount of items at or very near his or her level of ability. Thus, the precision of IRT ability scores for a given test varies across the band of test scores, with the most precise scores near the mean of all scores and the least precise at the extremes (assuming all examinees have attempted all items).

The measure of the precision of an IRT score is the standard error of the estimate of the examinee's ability, which varies with examinee ability. The standard errors of the ability estimate (in terms of the CPT Scale score) for both the Pre-CPT and the CPT, rounded to the nearest whole score, are presented in Appendix E: Standard Error of the Estimate for the Pre-CPT and CPT Across CPT Scale Scores.

A traditional standard error of measurement (SEM) may be calculated on the CPT score scale by using the standard deviation of the LOGIT scores and the KR-20 reliabilities, converting the results to the CPT Scale score. Table 16 presents the results of this process for the Pre-CPT and the CPT, by subsection.

Table 16
Standard Error of Measurement (SEM)
for the Pre-CPT and CPT
(in CPT Scale Scores)

Section	Pre-CPT	CPT
Listening	6.12	5.93
Reading	5.86	5.16
Structure	8.28	7.70

Table 16 reveals that the SEM for each subsection of the CPT is slightly smaller than for the Pre-CPT. This is due to the larger number of items in each subsection of the CPT.

Either the SEM given in Table 16 or the standard error of the estimate from the table in Appendix E can be used to construct confidence intervals around Pre-CPT and CPT scaled scores. For example, on a re-test, examinees will score within plus or minus one standard error of their scores about 67% of the time. Thus, using the table in Appendix E, if an examinee receives a score of 72 in the Listening section of the Pre-CPT, we can say that there is a 67% chance that the examinee would score between 66 and 78 on a re-test.

A careful examination of the table in Appendix E reveals that the shorter Pre-CPT measures with slightly more precision than the CPT at the lower end of the CPT Scale scores. At higher ability levels, the CPT measures with greater precision and across a wider spectrum than the Pre-CPT. Above scores of 96 to 97, the measurement precision of the Pre-CPT rapidly diminishes; a similar diminution of precision occurs with the CPT above scores of 150.

8.3 Validity

Validity refers to the extent to which a test actually measures what it purports to measure. The Pre-CPT and CPT claim to measure an examinee's proficiency in understanding authentic spoken and written Chinese, and ability to deal with Chinese structure.

The validity of any test cannot be "proven;" it can only be established by the collection of evidence that the test is indeed measuring what it purports to measure. Some commonly accepted types of evidence include evidence for content, concurrent, and construct validity. Each type of validity, and evidence which supports it in the case of the Pre-CPT, are explained below. Content validity is based on test content. The Pre-CPT and CPT are intended to be measures of proficiency in dealing with every-day 'real-life' Chinese. Validity based on content thus entails an examination of the degree to which the tests, in their stimulus passages, sample from the corresponding language-use situations the examinee might be expected to encounter in real-life.

As described in Sections 2 and 3 of this report, and particularly in Section 1.3, stimulus passages for both listening and reading were drawn, to the greatest extent possible, from authentic language sources. Item developers searched Chinese language newspapers, magazines, journals, street signs, postage stamps, train schedules, etc. for sources for stimulus passages for readings. They also listened to and transcribed Chinese language news broadcasts, movies, announcements, etc., for sources for stimulus passages for listening. (A list of the main sources used for Pre-CPT passages appears in Appendix F: Sources for Pre-CPT Listening and Reading Passages.) Texts were modified only to the extent that they needed to be clarified when taken out of the larger context, or, in the case of a few items on the Pre-CPT, were simplified to be made appropriate to the low-level being tested. Questions were designed to check comprehension of the meaning of passages, and, for more difficult passages on the CPT, to check comprehension of opinions, attitudes or inferences contained in the passages.

For the Structure section, the Pre-CPT and the CPT test knowledge of correct Chinese syntax. Unlike listening or reading comprehension, this is not a real-life language-use task, though clearly knowledge of Chinese syntax is a part of comprehending spoken and written

Chinese. Passages used in this section on the Pre-CPT, though again based on authentic reading materials, are sometimes altered in order to meet the goal of testing knowledge of syntax. For both tests, in this section only Chinese is used for both the stimulus and response options.

In terms of content validity, then, it can be demonstrated that the Pre-CPT and CPT items have been drawn from real-life use of Chinese language.

A second type of evidence of validity is concurrent validity. Concurrent validity refers to the extent to which a test score correlates with results that may be obtained through the use of independent criteria external to the test, measured at the same point in time, to see if expected relationships exist.

For the Pre-CPT and the CPT, two external criteria may be used: the reported level of Chinese study and the home language used. It would be expected that scores on the tests would increase as amount of study increases, and that at the same level of study, native speakers of Mandarin would perform better in listening (though not necessarily in reading written Chinese or in structure) than speakers of other Chinese languages, and both perform better than those who do not speak any Chinese at home. Table 13 indicates that this is generally the case for the Pre-CPT and the CPT, with the exception of some means which are based on a very small number of examinees. Table 13 thus provides evidence of the concurrent validity of the two tests as measures of Chinese language proficiency.

A third way of examining validity is construct validity. The goal of construct validity is to determine whether or not a test measures a single underlying trait. One assumption of the Rasch model (and most IRT models) is that the items are "unidimensional;" that is, only one examinee ability or trait is necessary to account for performance on the test (Hambleton & Swaminathan, 1985, p. 16). The fit statistics provided by the Rasch model provide evidence of the extent to which unidimensionality exists. It may be argued that if the majority of items are appropriately fitting, then there is strong evidence for the construct validity of the test. Table 12, which indicates that only 8% or less of the items on both the Pre-CPT and the CPT in any section were misfitting using commonly accepted criteria, provides strong evidence for the construct validity of these measures.

8.4 Intercorrelations Among Test Subscores

The three sections of the Pre-CPT and CPT are designed to measure different skills within the general domain of Chinese proficiency. It is expected that these skills are interrelated; i.e., persons who are highly proficient in one skill area will tend to be proficient in the other areas as well. However, the intercorrelations are not expected to be perfect. If they were, there would be no need to report scores for each section; the subscores would represent the same rather than different aspects of language proficiency.

Table 17 reports the Pearson product-moment correlation coefficients measuring the extent of relationships among the three subsections for each test based on CPT Scale scores for over

2200 examinees who had taken the CPT prior to June, 1991, and the 651 examinees who participated in the norming administration of the Pre-CPT. The correlations have been disattenuated to account for errors of measurement.

Table 17
Intercorrelations Among Subscores

	Pre-CPT			CPT		
	List	Read	Str	List	Read	Str
Listening	----			----		
Reading	.68	----		.80	----	
Structure	.70	.88	----	.88	.87	----

Table 17 shows that though there is a fairly strong relationship among the skills tested by the three subsections of the test, each of the subsection scores provides some unique information about the examinee's proficiency in the Chinese language. The lower correlations between the Listening Comprehension section and the other sections for the Pre-CPT, when compared to the CPT, is most likely due to the relatively large number of native speakers of Chinese in that sample. It may be remembered that over 9% of the Pre-CPT sample received perfect scores in Listening Comprehension. Thus, for these examinees the Pre-CPT listening section exhibits a ceiling effect. This effect lowers its correlation with the other sections.

The lower correlation between the Listening section and the other sections is also due to the fact that the Reading and Structure sections require the examinee to read Chinese characters, whereas the Listening section involves spoken language only. Thus, the differences in the correlations support the interpretation that the Reading and Structure sections test understanding of written Chinese, while the Listening section tests understanding of spoken Chinese.

In summary, the patter of intercorrelations between the test scores supports the validity of the constructs the Pre-CPT and the CPT claim to measure.

This chapter has presented information on the psychometric properties of the Pre-CPT. CAL intends to report on future studies involving its Chinese language tests that may help further clarify their psychometric properties. If you have used these tests for research purposes, the staff at the Chinese Language Testing Program requests copies of any papers or reports stemming from that research.

REFERENCES

ACTFL. (1987). ACTFL Chinese proficiency guidelines. *Foreign Language Annals, 20,* 471-487.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinhart and Winston.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer-Nijhoff Publishing.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: SAGE Publications.

Henning, G. (1987). *A guide to language testing.* Cambridge, MA: Newbury House.

Stansfield, C.W. (Ed.) (1985). *Technology and language testing.* Washington, DC: Teachers of English to Speakers of Other Languages.

Wang, L.S., & Stansfield, C.W. (1988). *CPT test interpretation manual.* Washington, DC: Center for Applied Linguistics.

Wright, B.D., & Linacre, J.M. (1991a). *A user's guide to BIGSTEPS: Rasch-model computer program.* Chicago: MESA Press.

Wright, B.D., & Linacre, J.M. (1991b). *BIGSTEPS* [Computer Program]. Chicago: MESA Press.

Wright, B.D., & Masters, G.N. (1982). *Rating scale analysis.* Chicago: MESA Press.

Wright, B.D., & Stone, M.H. (1979). *Best test design.* Chicago: MESA Press.

APPENDIX A

Pre-CPT Field Test Participants

Pre-CPT Field Test Participants

Name of Institution	State	Number of Students
American University	DC	14
Foreign Service Institute	VA	13
George Washington University	DC	8
Georgetown University	DC	31
Montgomery College	MD	12
Ohio State University	OH	16
Stanford University	CA	11
University of Hawaii	HI	6
University of Pittsburgh	PA	8
University of Maryland	MD	11
University of Massachusetts	MA	75
11 universities/institutes		205
Bethesda Chevy Chase High School	MD	20
Richard Montgomery High School	MD	18
Northfield-Mount Herman High School	MA	31
Sidwell Friends High School	DC	10
Xaverian Brothers High School	MA	15
5 high schools		94

NOTE. These are the actual number of participating schools and students. Due to late administrations at two schools, some examinees could not be included in the data analysis. The database for the item analysis contained 272 examinees.

APPENDIX B

Examinee Background Questionnaire

Pre-CPT FIELD TESTING
BACKGROUND INFORMATION SHEET

Under SPECIAL CODES, fill in the answers to the following questions:

K: How many years of Chinese language study, NOT INCLUDING THE CURRENT YEAR, have you **completed**?

0 = None
1 = 1 year
2 = 2 years
3 = 3 years
4 = 4 or more years

L: How many college course credits, NOT INCLUDING YOUR CURRENT COURSE, have you **already earned** in Chinese?

0 = None
1 = 1 to 3
2 = 4 to 6
3 = 7 to 9
4 = 10 to 12
5 = 13 to 15
6 = 16 to 18
7 = 19 or more
8 = NOT APPLICABLE

M: Which of the following most appropriately describes the **level** of your CURRENT Chinese language class?

0 = NONE OF THE GIVEN OPTIONS
1 = Third Year High School
2 = Fourth Year High School
3 = Fifth Year High School
4 = First Semester College (First Year Chinese)
5 = Second Semester College (First Year Chinese)
6 = Third Semester College (Second Year Chinese)
7 = Fourth Semester College (Second Year Chinese)
8 = Fifth Semester College (Third Year Chinese)
9 = Sixth Semester College (Third Year Chinese)

N: For how many **hours a week** does your current Chinese language class meet?

0 = Not Applicable
1 = 1 hour/week
2 = 2 hours/week
3 = 3 hours/week
4 = 4 hours/week
5 = 5 hours/week
6 = 6 hours/week
7 = 7 hours/week
8 = 8 hours/week
9 = 9 or more hours/week

O: Are you of a Chinese ethnic heritage?

0 = Yes
1 = No

P: Do you speak Chinese at home?

0 = Yes, Mandarin
1 = Yes, but not Mandarin
2 = No

APPENDIX C

Pre-CPT Norming Administration Participants

Pre-CPT Norming Administration Participants

Names of Institution	State	Number of Students
California State University, LA	CA	8
Connecticut College	CT	6
Cornell University	NY	25
Harvard University	MA	46
John Hopkins University	DC	6
Middlebury College	VT	12
University of California, SD	CA	87
University of Hawaii	HI	32
University of Oregon	OR	26
University of North Carolina	NC	14
University of Iowa	IA	24
University of Minnesota	MI	8
University of Virginia	VA	4
Wellesley College	MA	11
14 Universities		309
Barstow High School	MO	9
Bronx High School	NY	29
George Washington High School	CA	29
Isidore High School	LA	9
Lowell High School	CA	123
Phillips Academy	MA	27
Ridgewood High School	NJ	19
Seaholm High School	WI	13
Shady Side High School	PA	8
Springfield High School	MA	17
St. Louis High School	MO	19
University School of Milwaukee	WI	11
12 High Schools		313
Chinese School of Delaware	DE	10
Potomac High School	MD	19
2 Weekend Schools		29

Note: These are the actual number of participating schools and students. Due to late administrations at some schools, some examinees could not be included in the data analysis. The database for the item analysis contained 651 examinees.

APPENDIX D

Norming Tables for the Pre-CPT and CPT

TABLE A
Pre-CPT
Percentile Rank Table
All Students

	High School			First Year University		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
above 150	99	99	99	99	99	99
150						
149						
148						
147						
146						
145						
144						
143						
142						
141						
140						
139						
138		97			98	
137			93			95
136						
135	92			92		
134	79					
133						
132						
131						
130			83			
129						
128						
127						
126		88			94	
125		83				
124						
123			82			88
122	79			82		78
121				71		
120					90	
119						
118		83			90	
117						
116						
115						
114	71		71	70	84	77
113						
112		78	63		83	
111						71
110						
109						
108	61	73	63	66	80	70
107					77	62
106		69			76	
105						
104	58	69		61	76	62
103					71	
102			57		70	61
101		64	50		70	

Pre-CPT
Percentile Rank Table
All Students

	High School			First Year University		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
100	54			54		53
99					65	
98		61	49	49	64	52
97					60	
96	51		43	48		
95		55			60	45
94		52				
93	46	52	43	46	55	44
92		48			51	
91	43	48		43	51	39
90			39			39
89		45	34		45	35
88	42	41		39	40	34
87				36	39	
86	40	41	34	36	39	34
85		39	30	33	34	30
84					32	30
83	38	36	30	32	31	30
82				29	27	25
81	36	33		29	27	
80		31				
79	34	31	26	27	23	24
78		28		25	20	19
77	31	28		25	20	19
76	29	25	23	21	18	18
75			20		15	
74	27	23		19	15	15
73			20			15
72	25	20		16	14	
71		18			12	
70	24		16	12	11	10
69	23	17		11	10	
68	21	14	15		9	
67	20		14	9		
66	19	12	14		7	8
65	19			8		
64	18	10		6	6	
63		9	10			7
62	17					
61	14	8		5	5	5
60						
59	12		7	3		4
58		7		2	4	
57	11			2		
56	9	6	6	1	3	
55						
54	7	6			2	
53	6					2
52	5	5			1	
51	4	4	5	1	1	
50						
below 50	3	2	3	0	0	2

TABLE B

Pre-CPT

Percentile Rank Table

English-Speaking Students Only

	High School			First Year University		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
above 150	99	99	99	99	99	99
150			99			
149						
148						
147						
146						
145						
144						
143						
142						
141						
140						
139						
138		99			99	
137			99			97
136						
135	98			98		
134						
133						
132						
131						
130						
129						
128						
127						
126		98			96	
125						
124						
123			96			96
122	93			96		
121						
120					93	
119						
118		97			92	
117						
116						
115						
114	90		92	88	89	89
113						
112		97	87		88	
111						
110						
109						
108	85	96	86	85	86	82
107					85	76
106					84	
105						
104	81	92		82	83	
103						
102			81			76
101		90	72		79	

Pre-CPT
Percentile Rank Table
English-Speaking Students Only

	High School			First Year University		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
100	78			76		66
99					76	
98		87	71	71	75	65
97						
96	76			71		
95		83			72	
94		80				
93	73	80	66	68	64	56
92		74				
91	69	73		63	60	49
90			61			48
89		70	55		55	44
88	68			58	49	43
87				54		
86	65	64	55	54	48	42
85		60	50	50	43	39
84					40	38
83	63	55	49	49	38	37
82				46	32	31
81	57	49		46	31	
80		48				
79	55	47	44	43	28	30
78				40		24
77	49	44		40	24	24
76	47	40	39	34	22	23
75			34			
74	45	37		31	18	19
73			33			19
72	40	32		27	17	
71		28			16	
70			26	19	14	12
69	39	26		17	13	
68	34	22	24		10	
67	33			14		
66	31	18	23		8	11
65	30			12		
64	27	15		9	7	
63		14	18			9
62	25					
61	23	13		6	6	6
60						
59	19		12	5		6
58		12		3	4	
57	16			3		
56	14	11	11	2	3	
55						
54	10	10			1	
53	8					2
52	7	8				
51	5	6	8	1		
50			6			
below 50	4	5	5	1	1	1

TABLE C
CPT
Percentile Rank Table

	Beginning			Intermediate			Advanced		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT	LIST	READ	STRUCT
Above 164	99	99	99	99	99	99	99	99	99
164						99			98
163								99	
162									
161				99			99	98	
160									
159									
158									
157									
156							99		
155		99	99			99		98	96
154									
153									
152									
151	99						98		
150							97		
149		99			99	99		95	95
148									
147	99			99			97		
146									
145		99		99	99			92	
144	99			99			97		
143		99	99			99			93
142						99			
141		99		99	98			89	
140	99			99			96		
139			99			99			90
138	98			98			94		
137		98			98			86	
136									
135	98		99	98		98	92	83	88
134		98			98			83	
133	97			97			90		
132							88	80	
131		97	98		96	97		79	86
130	97			97		97	87	77	
129		97			95	97	86	77	
128	96			96			85	73	
127	95					97		73	82
126	95	97		95	95	96	82	73	
125					93	96		70	79
124	93	96	98	93	93	96	80	69	79
123					92	94		67	
122	92	96		93	91	94	77	67	
121	91		97	92		94	73	63	77
120					90	91		63	72
119	91			90	88	91	69	61	72
118		96	97		88	91	66	60	72
117	88			89	86		66	59	65
116	88	95			86	89	62	58	65
115	88	95	97	87	84	89	62	53	65

CPT
Percentile Rank Table

	Beginning			Intermediate			Advanced		
Scale Score	LIST	READ	STRUCT	LIST	READ	STRUCT	LIST	READ	STRUCT
114	86		96	85	82	86	59	50	60
113		93	96	83	82	86	55	50	60
112	84	93		83	80	82	55	46	54
111		93		81	80	82		46	54
110	83	92	95	81	78	81	51	43	53
109	83	91	94	78	76	76	46	40	
108		91	94	75	75	76		39	
107	83	90	93	75	73	76	42		47
106	82	90	87	72	73	71	39	36	41
105	80	89	87		70	71	35	33	41
104	80			68	67	65	35	31	35
103	78	88	84	65	66	65	31	30	35
102		87	83	62	64	64	27	27	35
101	77	86		61	61	59	27	24	28
100	77	85	79	58	60	59	24	23	
99	76	82	78	58	56	59	24	20	28
98	72	80	73	53	52	52	21	17	22
97		79	73	50	52	52	19	16	21
96	69	76	68	50	48	45	18	14	18
95	66		67	46	46	44	16	13	17
94	62	73	67	43	45	44	14	13	17
93	61	69	60	43	40	38	13	11	13
92	58	63	60	37	37	37	11	9	12
91		63	59	33	36	37	10	9	12
90	54	57	51	33	31	31	9	7	9
89		57	51	29	31	30	9	6	9
88	49	50	43	29	26		8	6	
87	44	44	42	26	23	25	6	4	6
86	40	43	41	22	22	25		4	6
85	40	35	32	22	18	19	5	3	4
84	34	31	31	17	14	19			3
83	34	30	30	17	14	18	4	2	4
82	28	25		14	11	13	3	2	3
81	20	25	20	11	11	13	2	1	
80	19	20	20	11	9	12	2	1	
79	17	19	19	8	8	12			3
78	16	15	14	8	6	9	1		2
77	11	11	14		4	9			
76	11	11	13	6	4	8	1		2
75				4		6			
74	8	8	7	4	3	5	1		1
73	5		7	3	2	5			1
72	5	5		3		3			
71		5	5		2	3			
70	2			2		3			
69	1	3	5		2	3			
68		2				2			
67		2	3	1	1	1			
66	1				1				
below 66	1	1	2	1	1	1	1	1	

APPENDIX E

Standard Error of the Estimate for the Pre-CPT and CPT Across CPT Scale Scores

TABLE D

Standard Error of the Estimate for the Pre-CPT and CPT Across CPT Scale Scores

	Pre-CPT			CPT		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
164				12	13	14
163				11	13	14
162				11	13	14
161				11	12	13
160				11	12	13
159				10	12	13
158				10	12	12
157				10	11	12
156				10	11	12
155				9	11	12
154				9	11	12
153				9	10	11
152				9	10	11
151				9	10	11
150				8	10	11
149				8	9	11
148				8	9	10
147				8	9	10
146				8	9	10
145			23	8	9	10
144			23	8	9	10
143			22	8	8	10
142			22	7	8	9
141			21	7	8	9
140			21	7	8	9
139	20		20	7	8	9
138	20	20	20	7	8	9
137	19	20	19	7	7	9
136	19	19	19	7	7	9
135	18	19	18	7	7	9
134	18	18	18	7	7	8
133	17	18	17	7	7	8
132	17	17	17	6	7	8
131	16	17	17	6	7	8
130	16	16	16	6	7	8
129	16	16	16	6	7	8
128	15	15	16	6	6	8
127	15	15	15	6	6	8
126	14	15	15	6	6	8
125	14	14	15	6	6	8
124	14	14	14	6	6	8
123	13	14	14	6	6	8
122	13	13	14	6	6	8
121	13	13	13	6	6	7
120	12	13	13	6	6	7
119	12	12	13	6	6	7
118	12	12	13	6	6	7
117	12	12	12	5	6	7
116	11	11	12	5	6	7
115	11	11	12	5	6	7
114	11	11	12	5	6	7
113	11	11	11	5	5	7
112	10	10	11	5	5	7
111	10	10	11	5	5	7
110	10	10	11	5	5	7
109	10	10	11	5	5	7
108	9	9	11	5	5	7

Standard Error of the Estimate for the Pre-CPT and CPT Across CPT Scale Scores

	Pre-CPT			CPT		
Scaled Score	LIST	READ	STRUCT	LIST	READ	STRUCT
107	9	9	11	5	5	7
106	9	9	10	5	5	7
105	9	9	10	5	5	7
104	9	9	10	5	5	7
103	9	8	10	5	5	7
102	8	8	10	5	5	7
101	8	8	9	5	5	7
100	8	8	9	5	5	7
99	8	8	9	5	5	7
98	8	8	9	5	5	7
97	8	7	9	5	5	7
96	8	7	9	5	5	7
95	7	7	9	5	5	7
94	7	7	9	5	5	7
93	7	6	9	5	5	7
92	7	6	8	5	5	7
91	7	6	8	5	5	7
90	7	6	8	5	5	7
89	7	6	8	5	5	7
88	7	6	8	5	5	7
87	6	6	8	5	5	7
86	6	6	8	5	5	7
85	6	6	8	6	5	7
84	6	6	8	6	5	7
83	6	6	8	6	5	8
82	6	6	8	6	6	8
81	6	6	8	6	6	8
80	6	6	8	6	6	8
79	6	6	8	6	6	8
78	6	5	8	6	6	8
77	6	5	8	6	6	8
76	6	5	8	6	6	8
75	6	5	8	6	6	8
74	6	5	8	6	6	8
73	6	5	8	6	6	8
72	6	5	8	6	6	8
71	6	5	8	6	6	9
70	5	5	8	6	6	9
69	5	5	8	6	6	9
68	5	5	8	6	7	9
67	5	5	8	6	7	9
66	5	5	8	7	7	9
65	5	5	8			
64	5	5	8			
63	5	5	8			
62	5	5	8			
61	5	5	8			
60	5	5	8			
59	5	5	8			
58	5	5	8			
57	5	5	8			
56	5	6	8			
55	5	6	9			
54	5	6	9			
53	5	6	9			
52	5	6	9			
51	5	6	9			
50	5	6	9			

APPENDIX F

Sources for Pre-CPT Listening and Reading Passages

Pre-CPT Sources of Materials: some samples

Periodicals:

宇宙光	Yuzouguang
中国风土趣谈	Traditions and Customs in China
思想品德	On Personal Virtues
茶话与茶经	On Tea
中国科学院院刊	Journal of Chinese Academy of Science
神洲学人	Journal of Overseas Chinese Students
现代汉语	Chinese Language Today
华语世界	The World of Chinese Language

Novels:

山坳上的中国	China in a Small Mountain Valley
月亮的女儿	Daughter of the Moon
撒哈拉沙漠的故事	Stories from Sahara Desert

Movie Scripts:

婚礼	The Wedding
陌生人	A Stranger
邻居	Neighbours
马生	Mashen

Newspapers:

新闻自由导报	Chinese Students Paper
人民日报	People's Daily

Other Materials:

video movies
conversation recordings
TV programs
radio programs

brochures
diploma
identification card
recipes
medicine labels